# Στατιστική στην πληροφορική

## 1η Άσκηση

**a)**

**stemplot:**

**Δεδομένα I**

| | |
|---|---|
| 30 | 3 |
| 31 | 01 |
| 32 | 167 |
| 33 | 46 |
| 34 | 25 |

**Δεδομένα II**

| | |
|---|---|
| 0 | 0028 |
| 1 | 24 |
| 3 | 2 |
| 4 | 2 |
| 6 | 4 |
| 9 | 0 |

**Δεδομένα III**

| | |
|---|---|
| 0 | 0168 |
| 1 | 03567788 |
| 2 | 00156 |
| 3 | 059 |
| 4 | 013468 |
| 5 | 24899 |
| 6 | 06 |
| 8 | 16789 |
| 9 | 46 |

**boxplot:**

**Δεδομένα I**

**Max = 34.5**

**Min = 30.3**

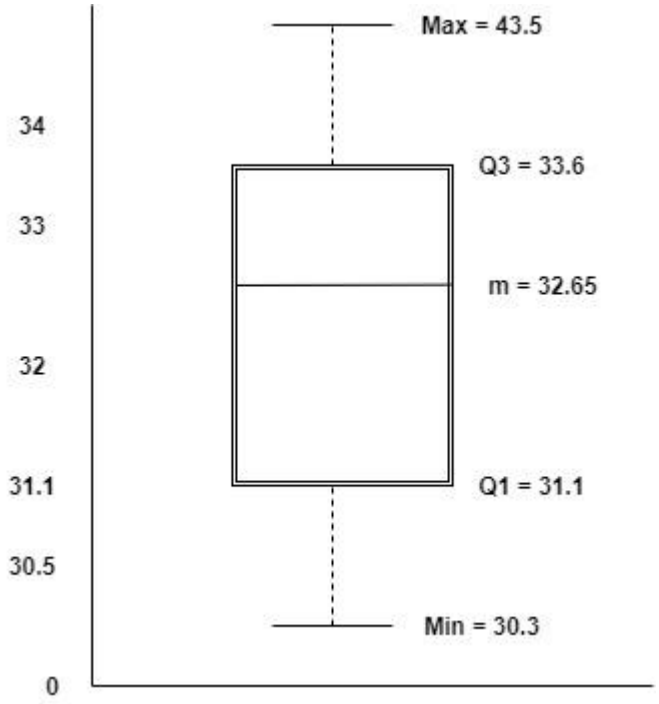
**Median = 32.65**

**Q1 = 31.1**

**Q3 = 33.6**

**Maximum = 37.35**

**Minimum = 27.35**




**Δεδομένα II**

**Max = 9.0**

**Min = 0.0**

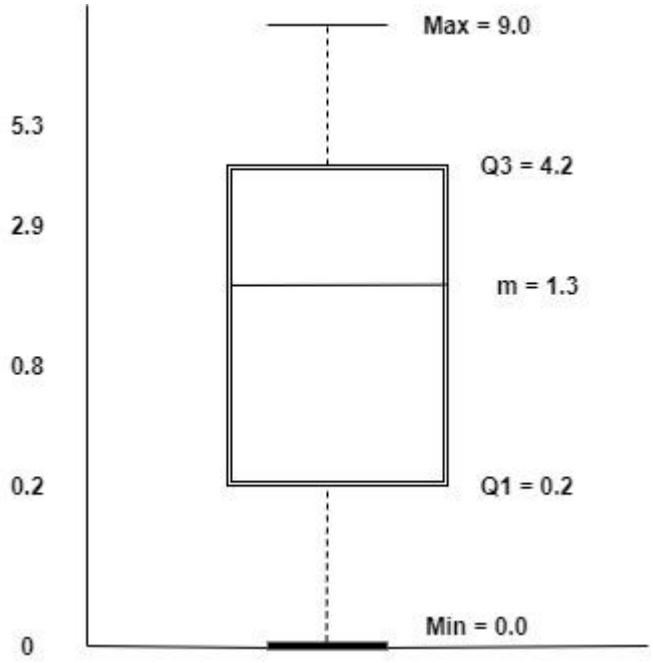
**Median = 1.3**

**Q1 = 0.2**

**Q3 = 4.2**

**Maximum = 10.2**

**Minimum = -5.8**




**Δεδομένα III**

**Max = 96**

**Min = 0**

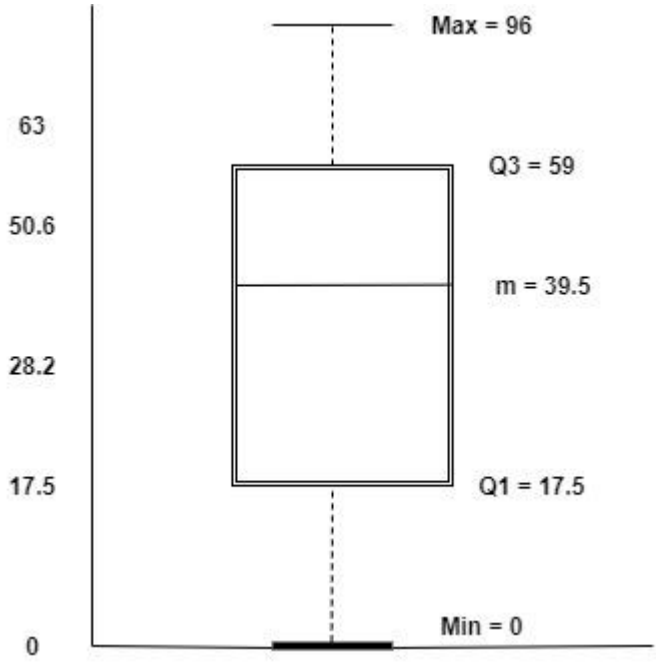
**Median = 39.5**

**Q1 = 17.5**

**Q3 = 59**

**Maximum = 121.25**

**Minimum = -44.75**




**b)**

**1.** Μέση τιμή = 32.55

Τυπική απόκλιση = 1.42

Median = 32.65

IQR = Q3 – Q1 = 2.5

Επειδή η μέση τιμή και η διάμεση τιμή δεν διαφέρουν σημαντικά, τότε μπορούμε να χρησιμοποιήσουμε και τους δύο τρόπους, όμως δεδομένου ότι η τυπική απόκλιση είναι μικρή, επιλέγουμε να χρησιμοποιήσουμε για την καλύτερη κατανομή των δεδομένων μας την μέση τιμή και την τυπική απόκλιση.

**2.** Μέση τιμή = 2.64

Τυπική απόκλιση = 3.06

Median = 1.3

IQR = Q3 – Q1 = 3.0

Θα προτιμήσουμε την σύνοψη των 5 αριθμών, διότι είναι πιο συμμετρικά κατανεμημένα τα δεδομένα, το οποίο προκαλείται κυρίως από μεγάλες τιμές όπως το 6.4 και το 9.0.

**3.** Μέση τιμή = 41.15

Τυπική απόκλιση = 28.26

Median = 39.5

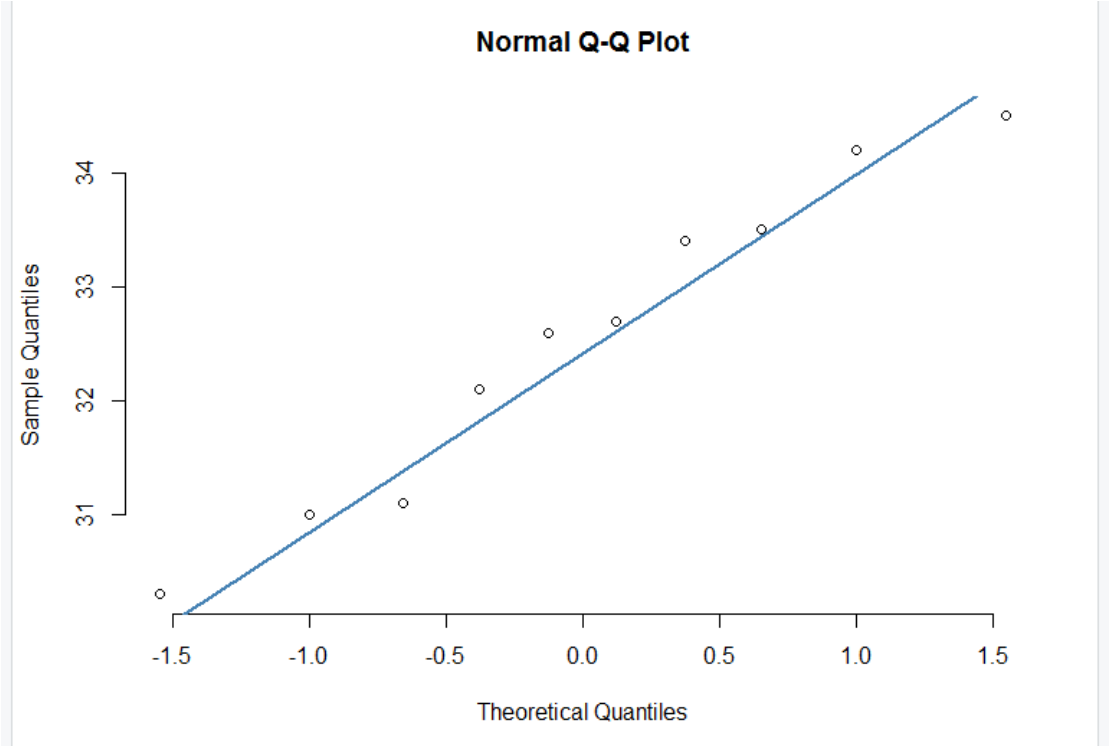
IQR = Q3 – Q1 = 41.5

Μπορούμε να χρησιμοποιήσουμε και τους δύο τρόπους, διότι η μέση τιμή και η διάμεση τιμή είναι πολύ κοντά μεταξύ τους, αλλά θα προτιμήσουμε την σύνοψη των 5 αριθμών γιατί είναι ελάχιστα καλύτερη η συμμετρικότητα των δεδομένων.

**c)**

- **Για τα ΔΕΔΟΜΕΝΑ Ι**

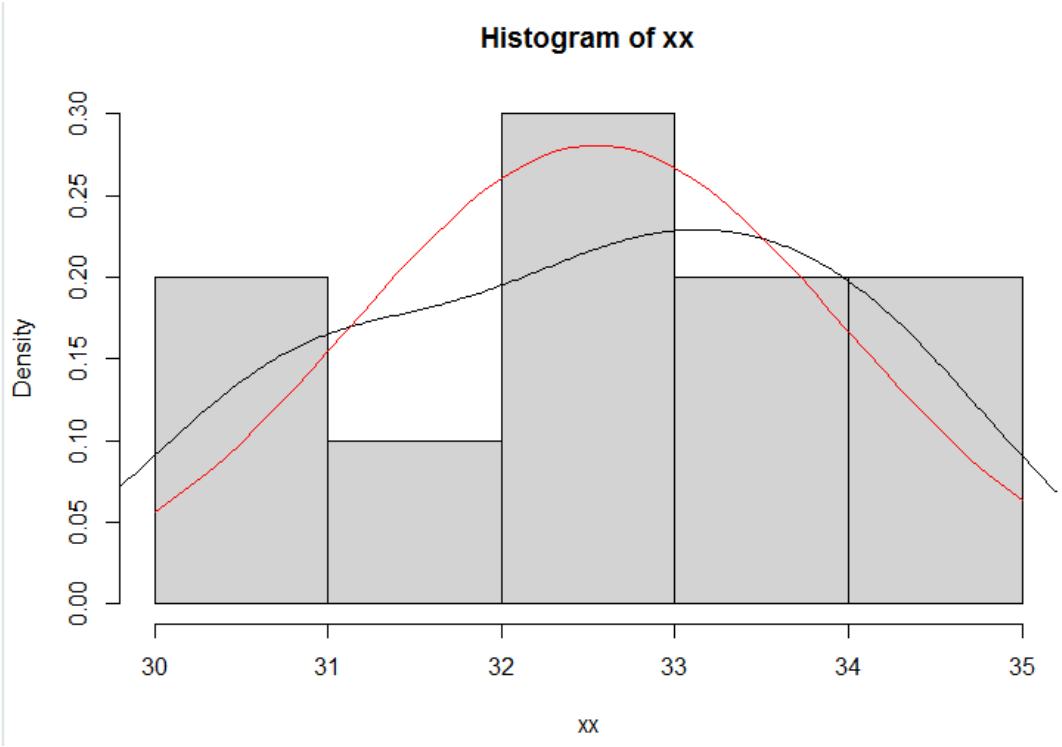
Αρχικά, κάνω το normal quantile plot στην γλώσσα R με το παρακάτω αποτέλεσμα:




Παρατηρώ ότι τα δεδομένα ακολουθούν την κανονική κατανομή, καθώς παρατηρούνται λίγες και μικρές αποκλίσεις από την ευθεία της κανονικής μορφής. Άρα θα ήταν ακριβής η προσέγγιση της κατανομής των δεδομένων από μια καμπύλη πυκνότητας της Κανονικής κατανομής.

Στη συνέχεια κατασκευάζω το ιστόγραμμα και τις καμπύλες πυκνότητας των δεδομένων και της αντίστοιχης κανονικής κατανομής, που έχει μέση τιμή και τυπική απόκλιση ίσες με αυτές των δεδομένων. Η καμπύλη φτιάχτηκε με τις εξής εντολές:

```
xx<-c(30.3,31,31.1,32.1,32.6,32.7,33.4,33.6,34.2,34.5)

hist(xx,probability = TRUE)

density(xx) -> pdf

lines(pdf)

x<- seq(30,35,0.1)

dnorm(x,mean=m,sd=sd) ->y

lines(x,y,col='red')
```

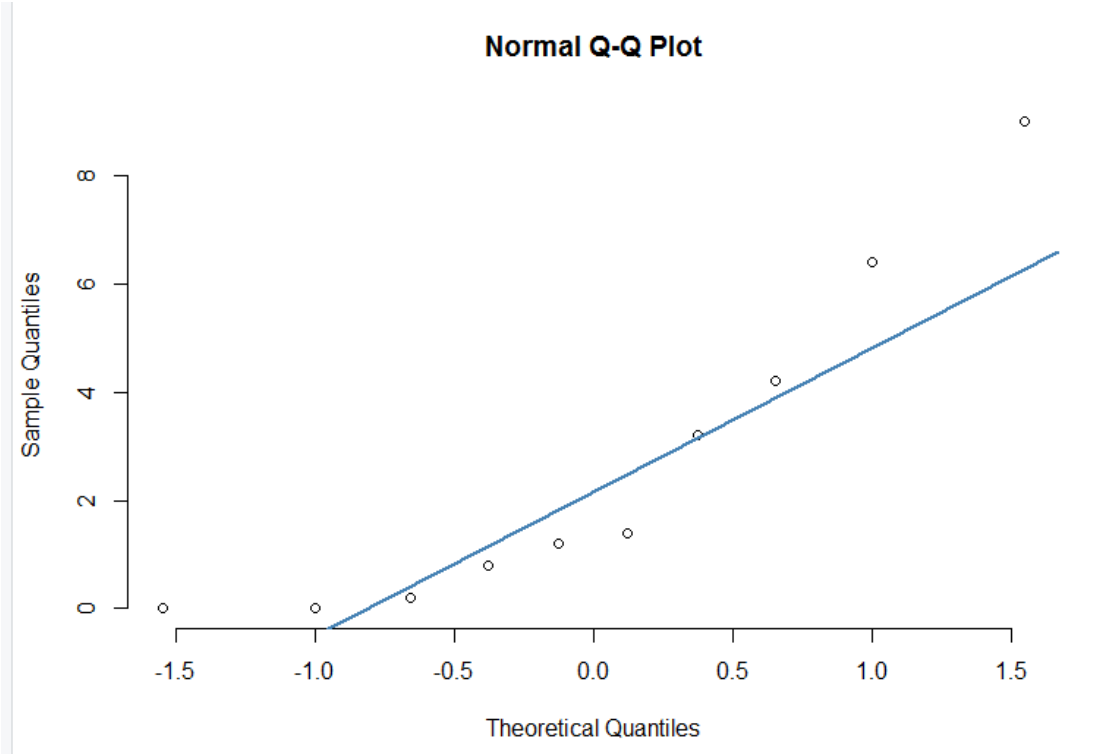



Για την εκτίμηση των αποκλίσεων θα κάνω το εξής: Έστω ότι θέλω να υπολογίσω την απόκλιση στο 25%. Αρχικά βρίσκω την τιμή σε αυτό το ποσοστημόριο για την κανονική κατανομή, με την εντολή << qnorm(0.25,mean=m,sd=sd) >>. Έπειτα υπολογίζω την αντίστοιχη τιμή για την κατανομή των δεδομένων με την εξής εντολή <<quantile(xx)>>.

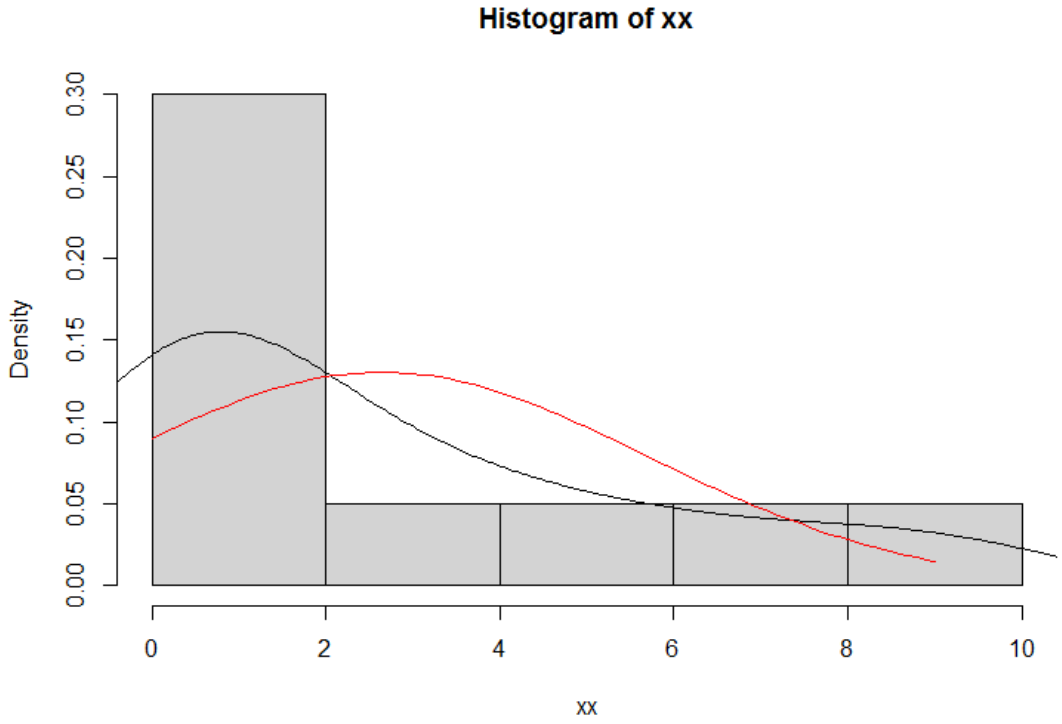
Παρατηρώ ότι η πρώτη εντολή βγάζει αποτέλεσμα 30.73, ενώ η δεύτερη 31.35. Αυτό είναι και το παράδειγμα απόκλισης η οποία εδώ είναι της τάξης του 2%. Η προσέγγιση είναι πολύ καλή.

- **Για τα ΔΕΔΟΜΕΝΑ ΙΙ**

Αρχικά, κάνω το normal quantile plot στην γλώσσα R με το παρακάτω αποτέλεσμα:




Στη συνέχεια κατασκευάζω το ιστόγραμμα και τις καμπύλες πυκνότητας των δεδομένων και της αντίστοιχης κανονικής κατανομής, που έχει μέση τιμή και τυπική απόκλιση ίσες με αυτές των δεδομένων. Οι εντολές είναι αντίστοιχα με την προηγούμενη περίπτωση.

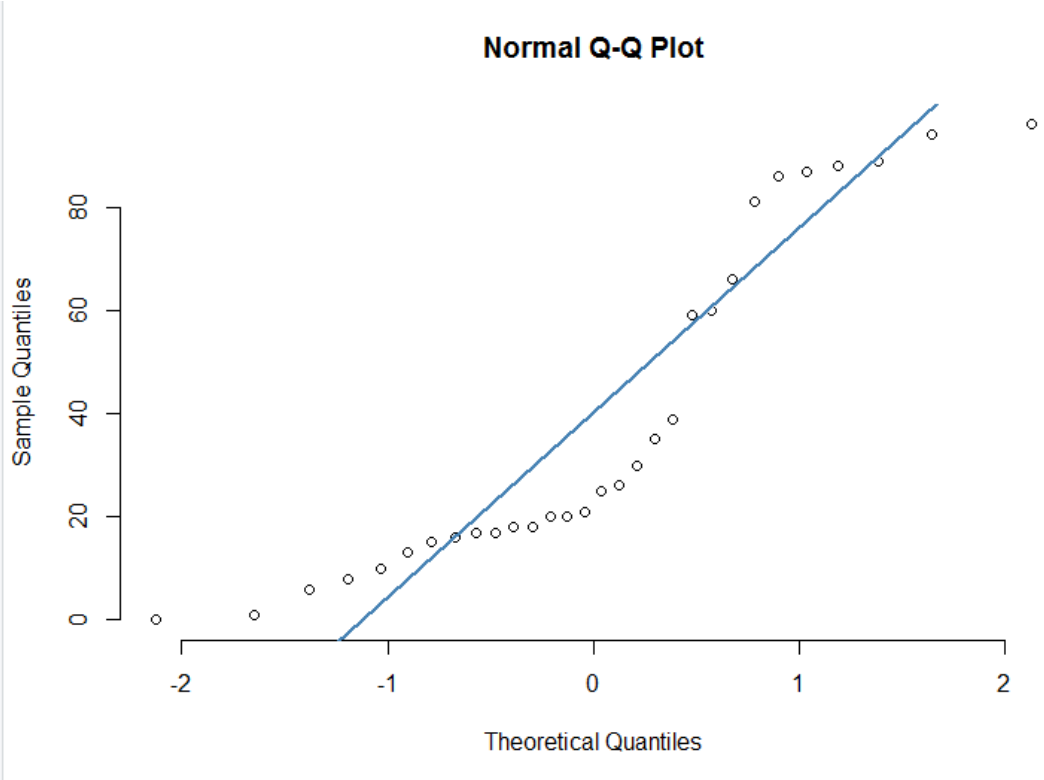



Για την εκτίμηση των αποκλίσεων θα κάνω το εξής: Έστω ότι θέλω να υπολογίσω την απόκλιση στο 25%. Η διαδικασία είναι ίδια με προηγουμένως.

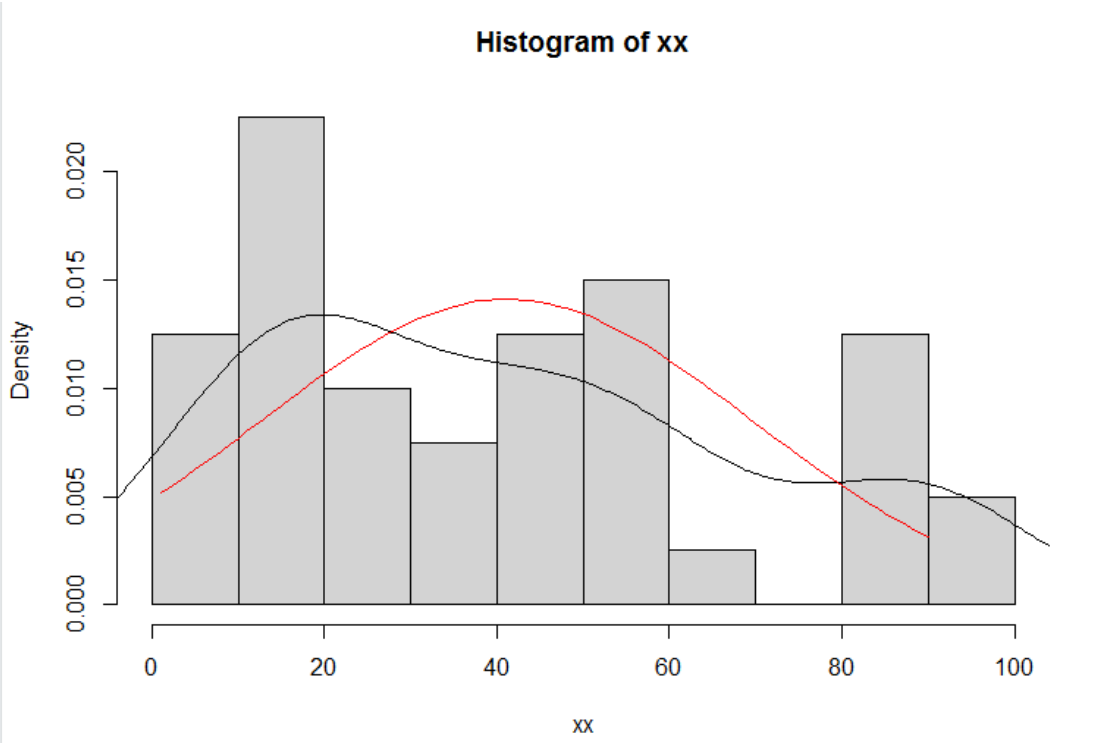
Παρατηρώ ότι η πρώτη εντολή βγάζει αποτέλεσμα 0.576 (δηλαδή το αποτέσματα της κανονικής κατανομής) ενώ η δεύτερη 0.35 (των δεδομένων). Στη περίπτωση αυτή, έχουμε μία απόκλιση της τάξης του 39% και προφανώς δεν είναι ακριβής η προσέγγιση.

- **Για τα ΔΕΔΟΜΕΝΑ ΙΙΙ**

Αρχικά, κάνω το normal quantile plot στην γλώσσα R με το παρακάτω αποτέλεσμα:




Στη συνέχεια κατασκευάζω το ιστόγραμμα και τις καμπύλες πυκνότητας των δεδομένων και της αντίστοιχης κανονικής κατανομής, που έχει μέση τιμή και τυπική απόκλιση ίσες με αυτές των δεδομένων. Οι εντολές είναι αντίστοιχα με την προηγούμενη περίπτωση.




Για την εκτίμηση των αποκλίσεων θα κάνω το εξής: Έστω ότι θέλω να υπολογίσω την απόκλιση στο 25%. Η διαδικασία είναι ίδια με προηγουμένως.

Παρατηρώ ότι η πρώτη εντολή βγάζει αποτέλεσμα 22.083 (δηλαδή το αποτέσματα της κανονικής κατανομής) ενώ η δεύτερη 17.75 (των δεδομένων). Στη περίπτωση αυτή, έχουμε μία απόκλιση της τάξης του 20% και η προσέγγιση λοιπόν είναι μέτρια.

**2η Άσκηση**

**a)** Τα στατιστικά δεδομένα προέρχονται από το επίσημο website της UIS (UNESCO Institute for Statistics). (http://data.uis.unesco.org/Index.aspx). Επιλέξαμε να πάρουμε δεδομένα που δείχνουν τον αριθμό των αμόρφωτων ανθρώπων ηλικίας από 15 έως 24 χρονών, που υπάρχουν παγκοσμίως από το 2014 έως το 2018. Όπως αντιλαμβανόμαστε, οι περιπτώσεις είναι όσες και οι χώρες που περιέχονται στον πίνακα, όμως εμείς επιλέξαμε δύο από αυτές τις χώρες: Μεξικό και Βραζιλία, επομένως έχουμε δύο περιπτώσεις.

**b)** Τα δεδομένα μας αποτελούνται από 3 columns. Οι δύο είναι κατηγορικές και η μία ποσοτική.

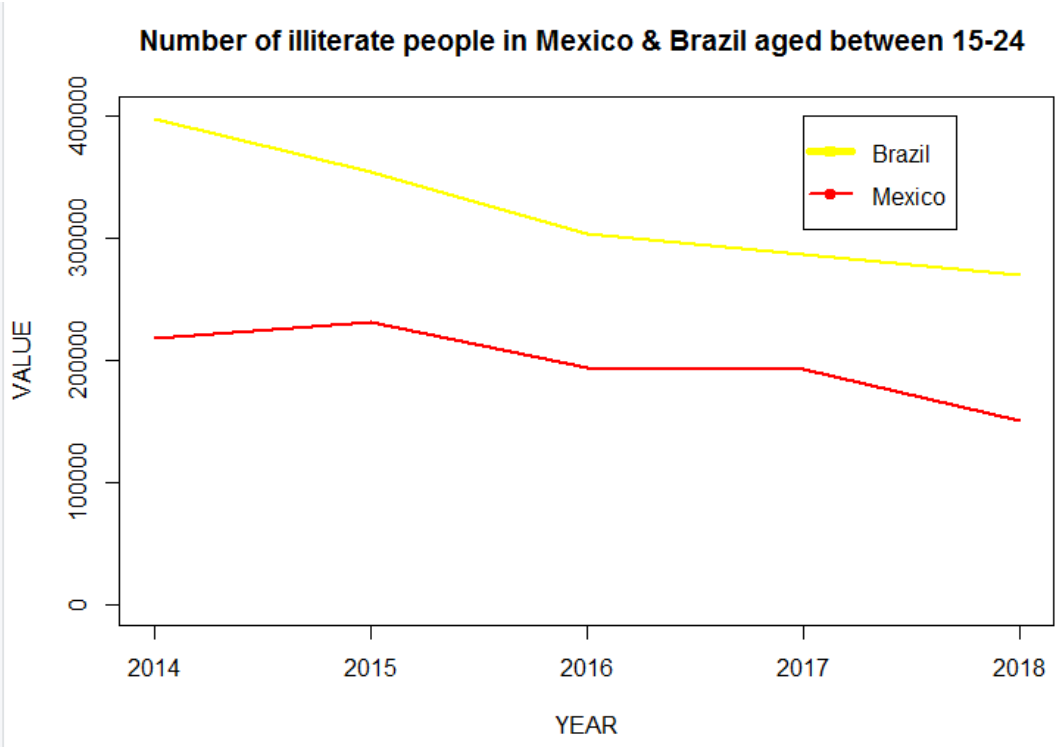
Χρόνος <- Κατηγορική μεταβλητή (σε ορισμένες περιπτώσεις μπορεί να είναι ποσοτική αλλά στην περίπτωση μας παίρνουμε δεδομένα από το 2014 έως 2018 και κάθε χρονιά είναι μια ξεχωριστή κατηγορία)

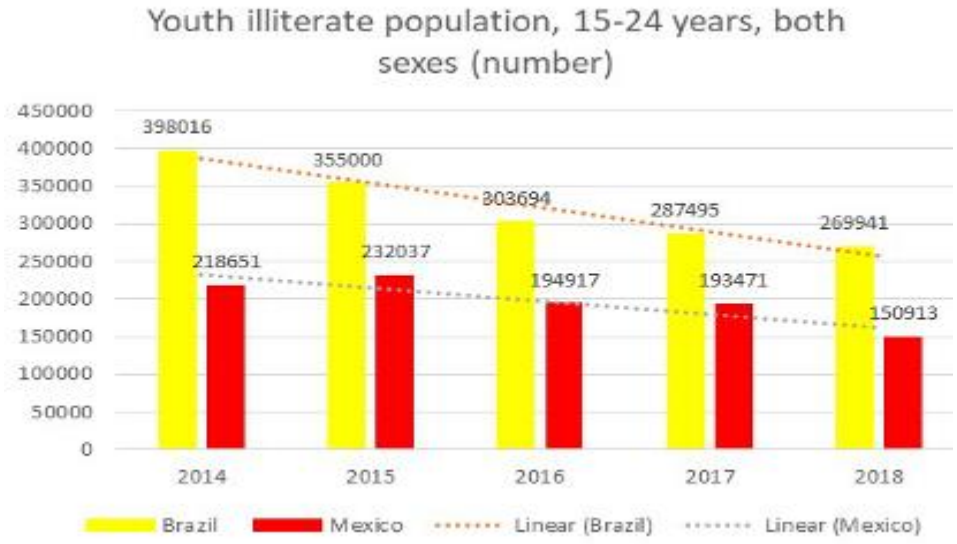
Αξία (ο αριθμός των ανθρώπων) <- Ποσοτική μεταβλητή

Χώρα <- Κατηγορική μεταβλητή

**c)**

Αρχικά ούτε στην κατανομή της Βραζιλίας, αλλά ούτε και του Μεξικού δεν παρατηρούμε outliers. Αυτό ίσως οφείλεται, διότι δεν παρατηρείται κάποια ριζική αλλαγή στον τρόπο διδασκαλίας στις χώρες αυτές, που να οδηγήσει σε μεγάλες αλλαγές ειδικά σε τόσο μικρό χρονικό διάστημα των 5 ετών. Η μορφή και των δύο κατανομών είναι φθίνουσα γραμμική, το οποίο μας δείχνει ότι με την πάροδο των χρόνων ολοένα και λιγότεροι νέοι άνθρωποι παραμένουν αμόρφωτοι. Ένας λόγος για τον οποίο μπορεί να συμβαίνει αυτό, είναι ότι με την εξέλιξη της τεχνολογίας, είναι πολύ πιο εύκολο να ψάξει και να βρει κανείς οποιαδήποτε πληροφορία χρειαστεί, με ελάχιστο κόπο και χρόνο. Αυτό καθιστά την μάθηση πιο εύκολη, μειώνοντας με αυτόν τον τρόπο τους αμόρφωτους ανθρώπους.




*<u>Η κατανομή των δεδομένων</u>*

**d)**

Υπολογίζουμε τυπική απόκλιση, μέση τιμή και σύνοψη 5 αριθμών και ξεχωριστά για κάθε χώρα.

**Μεξικό:**

**Μέση τιμή:**

mean(mexicodata$Value) [1] 197997.8

**Τυπική απόκλιση:**

sd(mexicodata$Value) [1] 30952.32

**Σύνοψη 5 αριθμών:**

summary(mexicodata$Value) Min. 1st Qu. Median Mean 3rd Qu. Max.

150913 193471 194917 197998 218651 232037

**Βραζιλία:**

**Μέση τιμή:**

mean(brazildata$Value) [1] 322829.2

**Τυπική απόκλιση:**

sd(brazildata$Value) [1] 52677.89

**Σύνοψη 5 αριθμών:**

summary(brazildata$Value) Min. 1st Qu. Median Mean 3rd Qu. Max.

269941 287495 303694 322829 355000 398016

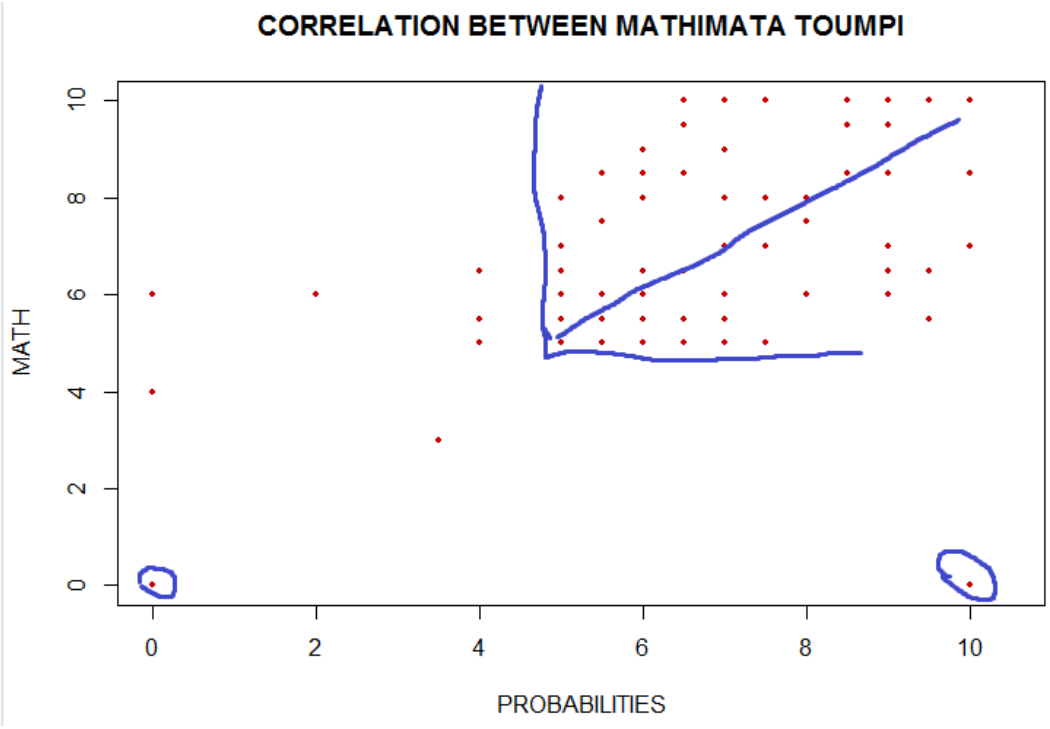
Παρατηρούμε ότι η μέση τιμή με την διάμεση δεν έχουν μεγάλη απόκλιση στις τιμές τους. Ένας λόγος που συμβαίνει αυτό είναι επειδή δεν έχουμε outliers. Συμπεραίνουμε λοιπόν ότι και οι δύο τρόποι είναι κατάλληλοι για να συνοψίσουμε την κατανομή, αλλά θα προτιμήσουμε να χρησιμοποιήσουμε την μέση τιμή και τυπική απόκλιση.

**e)**

Επιλέγουμε τις μεταβλητές χρονιά και value,όπου value ο αριθμός των αμόρφωτων ανθρώπων. Η σχέση αυτών των δύο μεταβλητών δεν είναι αιτιατή, καθώς οι αμόρφωτοι άνθρωποι επηρεάζονται και από άλλους παράγοντες, όπως η τεχνολογία και η οικονομική κατάσταση, παρ' όλα αυτά δεν μπορούμε να παραβλέψουμε ότι υπάρχει και μια σχέση μεταξύ τους.

**3η Άσκηση**

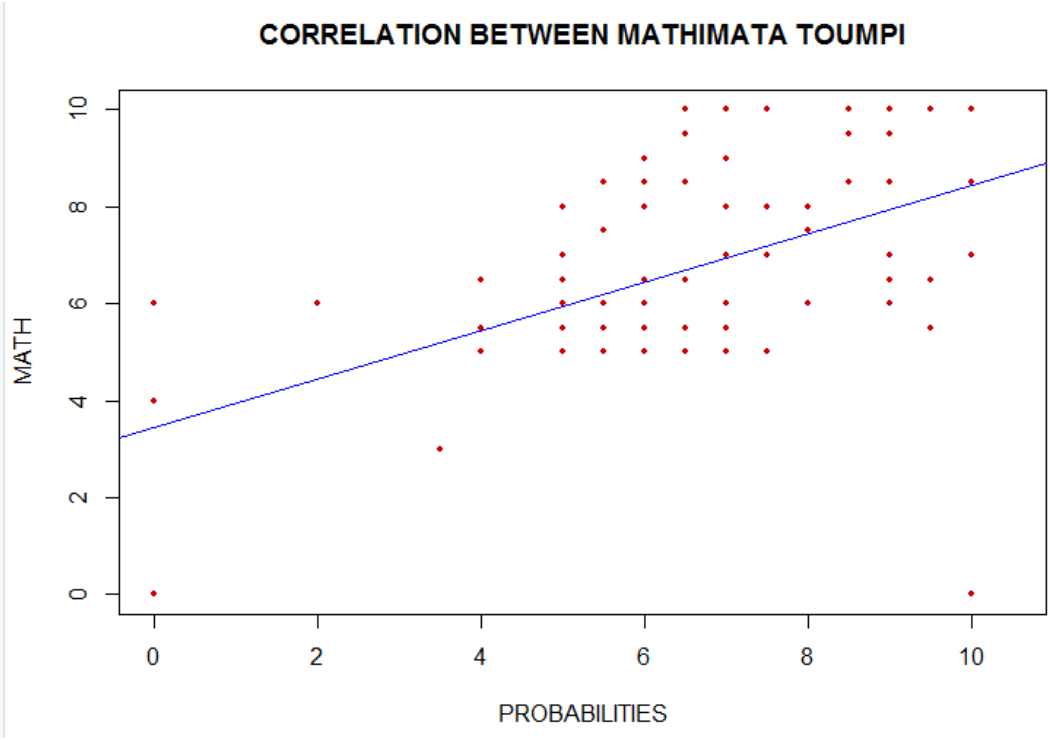
**a)**




Μέσω του scatterplot μοιάζει να έχει μορφή γραμμικής σχέσης, αλλά επειδή αποκλίνουν αρκετά ορισμένες τιμές, δεν είναι ξεκάθαρη αυτή η σχέση, ωστόσο θεωρούμε ότι υπάρχει. Επίσης έχει 2 outliers. Η κατεύθυνση της σχέσης δείχνει να κλείνει προς αύξουσα, αλλά το να πούμε απόλυτα κάτι τέτοιο δεν θα ήταν σωστό, οπότε καταλήγουμε πως παρουσιάζει μια ασθενή κλίση προς αύξουσα κατεύθυνση. Η δύναμη της σχέσης είναι ασθενής, διότι αν ήταν ισχυρή, θα έπρεπε όλες οι τιμές να είναι πάρα πολύ κοντά μεταξύ τους και να σχεδιάζουν μια γραμμή, αλλά στην συγκεκριμένη περίπτωση, δεν παρατηρείται απόλυτα κάτι τέτοιο γιατί οι τιμές είναι περισσότερο διασκορπισμένες μέσα στο χώρο.

**b)**

Με την βοήθεια της R και της συνάρτησης cor(x,y) όπου x = math και y = prob βρήκαμε ότι ο συντελεστής συσχέτισης είναι r = 0.5237964. Επίσης εκτελέσαμε την γραμμική παλινδρόμηση με την χρήση της εντολής abline(lm(y~x,data = survey_data_2020), col="blue") και το αποτέλεσμα ήταν το εξής:




**ΝΙΚΟΛΕΤΑ-ΚΛΕΙΩ ΠΑΤΑΤΣΗ p3180266**

**ΠΛΑΤΩΝΑΣ ΚΑΡΑΓΕΩΡΓΗΣ p3180068**